The Quarterly Journal of Experimental Psychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pqje20

The abstract representations in speech processing

Anne Cutler [a] [b]

[a] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[b] MARCS Auditory Laboratories, University of Western Sydney, Australia

Available online: 21 Oct 2008

To cite this article: Anne Cutler (2008): The abstract representations in speech processing, The Quarterly Journal of Experimental Psychology, 61:11, 1601-1619

To link to this article: http://dx.doi.org/10.1080/13803390802218542

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2008, 61 (11), 1601–1619

Psychology Press
Taylor & Francis Group

The 34th Sir Frederick Bartlett Lecture

The abstract representations in speech processing

Anne Cutler

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, and MARCS Auditory Laboratories, University of Western Sydney, Australia

Speech processing by human listeners derives meaning from acoustic input via intermediate steps involving abstract representations of what has been heard. Recent results from several lines of research are here brought together to shed light on the nature and role of these representations. In spoken-word recognition, representations of phonological form and of conceptual content are dissociable. This follows from the independence of patterns of priming for a word's form and its meaning. The nature of the phonological-form representations is determined not only by acoustic-phonetic input but also by other sources of information, including metalinguistic knowledge. This follows from evidence that listeners can store two forms as different without showing any evidence of being able to detect the difference in question when they listen to speech. The lexical representations are in turn

Correspondence should be addressed to Anne Cutler, Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands. E-mail: anne.cutler@mpi.nl

A version of this paper was presented as the 34th Bartlett Lecture to the Experimental Psychology Society in April 2006, under the title "Levels of Processing Speech". The findings described in the paper have arisen from a research programme of many years, and publications describing them have been coauthored with Dennis Norris, James McQueen, Andrea Weber, Takashi Otake, Sally Butterfield, and Frank Eisner. I am very grateful to all of these and particularly to the first two for, respectively, three and two decades of stimulating intellectual companionship. I am further grateful to them for helpful comments on an earlier version of this text, as also to the two reviewers of the manuscript, Gareth Gaskell and Jeff Bowers.

http://www.psypress.com/qjep
DOI:10.1080/13803390802218542

separate from prelexical representations, which are also abstract in nature. This follows from evidence that perceptual learning about speaker-specific phoneme realization, induced on the basis of a few words, generalizes across the whole lexicon to inform the recognition of all words containing the same phoneme. The efficiency of human speech processing has its basis in the rapid execution of operations over abstract representations.

Keywords: Speech processing; Phonemes; Lexicon; Representations.

Humans with unimpaired hearing extract meaningful information from acoustic input in countless ways. Their behaviour shows it—the leap out of the way on hearing a car approaching from behind, the sigh of relief as a letter falls into a still-full postbox, the grimace at the sound of breaking glass from the kitchen. Animals extract meaning from acoustic signals too, and it has recently become clear that their processing can extend beyond a simple associative mapping, such as from stimulus to threat or reward; animals of many species can derive meaning by inference beyond the actual form of an acoustic input (Arnold & Zuberbühler, 2006; Kaminski, Call, & Fischer, 2004). But, incontrovertibly, all of the above examples concern processing of a far simpler nature than communication by speech. Processing speech turns an acoustic signal into meaning via a number of intermediate steps, with intermediate representations at each level of processing. This paper describes some recent advances in our knowledge of the nature of and interrelationships between these representations.

It is an interesting time to be working on speech processing by human listeners. Research in this area has traditionally been divided between two disciplines. Speech perception has been studied by phoneticians, who determine how acoustic evidence in a speech signal motivates listeners' decisions about which speech sounds (phonemes) the signal contains. Word recognition and the comprehension of sentences and discourse have been the domain of psycholinguistics. Neither discipline strives for a complete model of speech processing. Psycholinguists largely take the "front end"—the initial processing applied to raw acoustic input—for granted, assuming that it will deliver a representation of the input that is in a form suitable for accessing stored lexical entities. In the 1970s, and to a lesser extent in the 1980s, psycholinguists energetically debated the "units of perception"—namely, the form of these representations that are suitable for accessing a lexicon. They never then debated about whether any such prelexical level existed. Now, however, they do, and the result has been an unprecedented convergence; the two traditionally separate domains of speech perception research have begun to overlap, since it is only by simultaneous attention to both domains that current theoretical issues can be decided. Thus, in recent years phoneticians have come to take word recognition into account, and psycholinguists have become more sensitive to the nature of speech. A central issue of joint attention has become the representations, both prelexical and lexical, involved in speech processing: whether these representations are abstract or are accumulated traces of episodes of experience; if they are abstract, how many different kinds of representation there are; and where (at which levels of processing) they play a perceptual role.

Questions of representation permeate the theorizing in speech perception research. An overview of the field's research programme would include: determining what processing is involved in understanding a spoken utterance, establishing what parts of this processing are specific to spoken language (but not other auditory signals such as the sound of the car or of the letter falling into the box), and then—vital for the psycholinguist—distinguishing the parts that are universal across languages versus the parts that are subject to language-specific influence. These questions concern the stages of processing and the nature of the processes at each stage. There are then also issues concerning the relationship between

stages: that is, the flow of information—whether processes at a given stage receive input from a single direction or bidirectionally, for instance. As in most areas of cognitive psychology, this last issue is a hotly debated one (see, for example, Norris, McQueen, & Cutler, 2000, and the commentaries it evoked; or McClelland, Mirman, & Holt, 2006, and McQueen, Norris, & Cutler, 2006b). In publications of the field these central issues—universality versus language specificity in the nature of the processes, and unidirectionality versus bidirectionality in the relationship between the processes—unquestionably play a major role. But the nature and form of the representations upon which speech perception processes operate also play a role in all these debates. Sometimes this role is implicit, but there is no way to take a theoretical stance without committing to a position on representational issues. No explanation without representation!

The following sections of this paper describe recent evidence that constrains the prelexical and lexical components of speech-processing models with respect to the nature and form of the representations involved. This evidence motivates the conclusions that spoken-word recognition involves separate lexical representations of word form and word meaning, and that the former (phonological representations) can be activated without necessary consequent activation of the latter (conceptual representations); that the phonological representations are not entirely and solely constrained by the output of prelexical processing; and that prelexical and lexical representations are separate and, in both cases, abstract.

Separate lexical representations of word form and word meaning

Words have meaning, which is arguably constant whether words are spoken, heard, or written. Across speaking, listening, and reading, however, the form of words differs—respectively, one and the same meaning is encoded by articulatory, acoustic, and orthographic forms. Models of lexical processing must stipulate the relationship between the representations of form and the representations of meaning. Psycholinguists who have built models of lexical access in speech production (e.g., Dell, 1988; Levelt, Roelofs, & Meyer, 1999) have found it necessary to formally separate the conceptual representations activated to express a chosen meaning from the phonological representations that contact the articulatory output mechanisms. Evidence for the separation of these two types of representation comes inter alia from frequency effects on speech production. If form and meaning representations were necessarily linked, then the frequency of the meaning should govern ease of access to the form. A hare and a swan, very similar in frequency, should be equally easy to name, and it should not matter that the name *hare* has a homophone *hair* with a much higher frequency. But it does matter; the naming time is determined by the highest frequency homophone (Dell, 1990; Jescheniak & Levelt, 1994).

In speech comprehension, evidence on this issue is harder to come by; a homophonous spoken form may be held to activate two lexical entries that are phonologically identical without this requiring separation between the phonological and conceptual components of each entry. Some models of lexical access in comprehension have indeed explicitly proposed that access to a word's form entails access to the word's meaning (e.g., Gaskell & Marslen-Wilson, 1997). In this section, however, evidence is presented from a series of experiments contrasting priming of semantically associated words with priming of phonological form; this evidence motivates the claim that the recognition of spoken words, like word production, draws on formally dissociable phonological and conceptual representations.

The experiments made use of two alternative versions of a single task: cross-modal priming. This task measures the effects of an immediately preceding heard prime on lexical decision responses to a visual target string. Presented with the letter string RIGHT, English-speakers should respond yes: It is a word. Do they do this faster after hearing the word *wrong* than after hearing an unrelated control word—for example, *soon*? If so, then we assume that processing of the lexical representation of *wrong* has facilitated

processing of the lexical representation of *right*—in this case because there is a relationship of meaning between the two words. If in an experiment the related primes and targets share such a relationship of meaning, we refer to associative priming. But primes and targets do not have to be related in meaning, because cross-modal priming also has an identity-priming variant. In identity priming, effects of the prime *wrong* might be assessed on lexical decisions to the target WRONG. It is clear that priming in this case should be very strong; reading a word just after hearing exactly the same word spoken should be much easier than reading it after hearing some other word.

These two versions of the cross-modal priming task allow us to use the task to probe the activation of the components of lexical representations. Briefly: If spoken *wrong* facilitates responses to WRONG, then at the very least the phonological form of the prime word has been activated. If spoken *wrong* facilitates responses to RIGHT, then at the very least the prime's conceptual representation has been activated. The relative activation patterns of the two types of representation can be compared, enabling researchers to ask such questions as whether activation of the one representation irrevocably entails activation of the other, or whether activation of the one representation necessarily precedes activation of the other.

Furthermore, we can ask these questions with respect to the simple relationship between the prime and target words (e.g., with the prime word presented in isolation), or with the prime word occurring in a wider sentence context, or even with the prime word occurring, as it were, by accident. Words occur by accident all the time. This is just an inevitable consequence of constructing a vocabulary of hundreds of thousands of words out of only a few dozen phonemes—the words cannot all be unique sequences, and especially long words will tend to contain short words as accidental embeddings (*you* in *unique*, *seek* in *sequence*, *axe* and *dent* and *dental* in *accidental*—every sentence contains a wealth of examples).

The issue for the spoken-word recognition researcher is the role of such accidentally present words in the recognition of the words that are intentionally present in the speech input. We are often not consciously aware of the accidental words, though sometimes our attention may be drawn to them. Some years ago, for instance, a *Cambridge Evening News* fashion feature sported the headline "Sarong—so right"; the writer who chose the headline capitalized on the fact that every time we hear the word *sarong* we hear within it all the phonemic evidence necessary to support the word *wrong*. Puns often exploit embedding in this way, and the fact that writers and speakers make such puns means that we are not always oblivious to the spurious occurrences of words such as *wrong* in *sarong*. But does hearing *sarong* necessarily activate the phonological representation of *wrong*? And does it, further, activate the conceptual representation of *wrong*?

In a series of experiments this string of related questions was put to the test by Norris, Cutler, McQueen, and Butterfield (2006b). Their study included all possible combinations of the two forms of the word (intended: *wrong* versus accidental: *sarong*), the two forms of the task (identity priming versus associate priming), and two types of context (none, i.e., isolated-word presentation, versus sentence contexts). Prime words such as *wrong* were thus presented in isolation, were presented accidentally embedded in carrier words (*sarong*), or were contained in sentence contexts, which again could be for either the word itself (*They were surprised to learn that the wrong costumes had been ordered*) or for its carrier word (*They were surprised to learn that sarong costumes had been ordered*). The effect of all of these different types of prime was measured on recognition of visually presented target words, which could be the word itself (WRONG) or a conceptually related associate (RIGHT).

The results clearly showed that phonological activation and conceptual activation can dissociate. In Norris et al.'s (2006b) experiments, *wrong* in isolation facilitated recognition of both WRONG and RIGHT, but *wrong* in a sentence

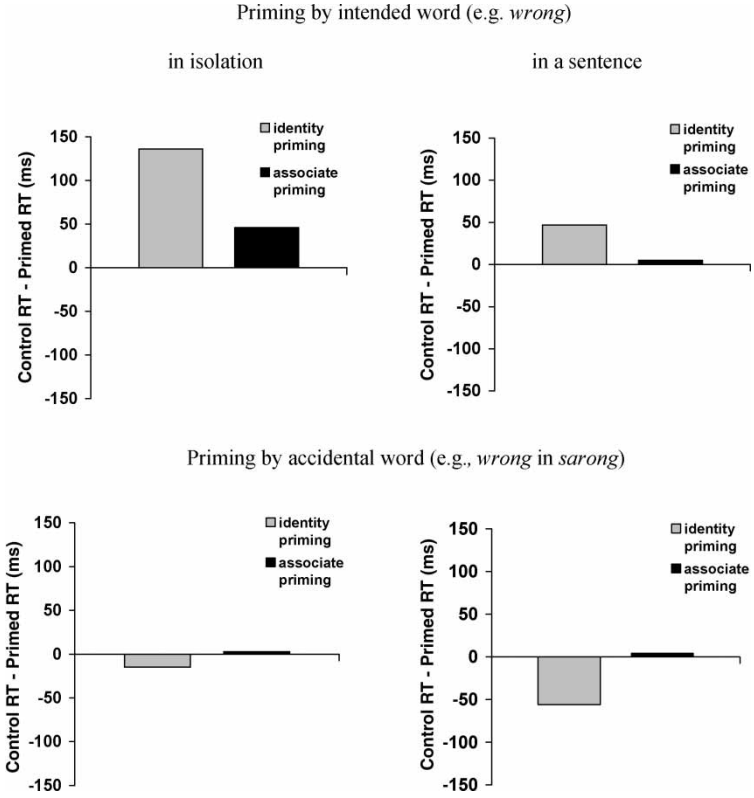

Figure 1. *Effects of cross-modal priming by spoken prime words such as* wrong*, expressed as difference in reaction time from a control condition with a different, unrelated prime, separately for visual target words such as WRONG (identity priming) and RIGHT (associate priming). In the upper figures, the prime was the intended word (*wrong*); in the lower figures, it was accidentally present (*wrong *in* sarong*); in the left figures, the prime was spoken in isolation (*wrong, sarong*); in the right figures, it was spoken in a sentence context. Based on results from Norris et al. (2006b).*

context facilitated only recognition of WRONG, not of RIGHT. The prime *sarong* never had any perceptible effect on recognition of RIGHT, but exercised inhibitory effects on recognition of WRONG, especially when it occurred in a sentence context. Figure 1 summarizes the results, expressed as priming effects, across the experiment series.

Thus phonological activation is robust. When we hear *wrong*, the appropriate phonological representation is activated. This is true when we hear *wrong* in isolation and when we hear *wrong* in a sentence, and it also occurs when we hear *wrong* accidentally embedded—for example, in *sarong*. In the latter case, activation is indirectly measurable via the consequent inhibition. In all well-known current models of spoken-word recognition (e.g., Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986; Norris, 1994), recognition is assumed to occur via a process of competition between multiple concurrently activated lexical candidates, which are fully or partially supported by the speech input; the more support a candidate word receives, the more it is able to inhibit competing words and thereby eventually win the competition. Inhibition is the sign that a candidate was activated, but lost out to a rival word in the competition process. When *sarong* is heard, *sarong* wins, but to win it has had to suppress activation of *wrong*, which was also competing.

In contrast to phonological activation, conceptual activation is less robust (or at least less independent). When we hear the word *wrong* uttered in isolation, its associated conceptual representation is strongly enough activated for an effect to be observed on recognition of the conceptual associate RIGHT. But when we hear *wrong* by accident, embedded in *sarong*, no detectable conceptual activation spreads to RIGHT. This holds when *sarong* occurs in a sentence and also when it is spoken in isolation. And even when we hear the word *wrong* itself in a sentence, there is not necessarily any conceptual activation either.

This last result was somewhat puzzling given that many previous studies, from Swinney (1979) onwards, had shown significant cross-modal associate priming from words in sentences. Norris et al. (2006b) further explored this aspect of their results and discovered that spoken *wrong* in a sentence produced significant priming for RIGHT if the sentence contexts were truncated immediately after the occurrence of the critical prime word (*They were surprised to learn that the wrong* –), or if the sentences contained contrastive accents (which effectively called up a discourse context in which the sentence could be placed). These conditions, or their equivalents, such as primes in sentence-final position, had been met in a large proportion of the published associate-priming studies with significant results. Thus the conceptual activation of an individual word is dependent on the conceptual activation of the whole context in which it occurs. A word in isolation may act as its own effective context, allowing activation of its full semantics, but a word in a sentence is constrained by the sentence context and the conceptual contribution it makes there.

Although these results thus also illuminate the relationship between word and sentence meaning, the principal finding for present purposes is the dissociation between the identity-priming and associate-priming situations. The implication of this dissociation is the separability of a word's phonological and conceptual representations. Effects arising in the identity-priming version of the cross-modal priming task reflect phonological activation consequent upon any occurrence of the word form, intended or not. Such effects have no necessary implications for conceptual activation (e.g., the rapidly inhibited activation of spuriously present embedded words appears to concern their phonological form alone). Effects in the associate-priming situation do not constitute a touchstone for activation of any kind, but are dependent on activation of lexical semantics, which may or may not accompany activation of phonological form. The one priming effect thus cannot be predicted from the other. A simple unitary model of lexical entries, in which phonological and conceptual representations are inextricably united, is therefore untenable; the recognition lexicon contains separate and independently functioning representations of form and of meaning.

Phonological representations and where they come from

Consider now the phonological representations, which, as Norris et al.'s cross-modal priming evidence attests, constitute the primary contact between speech input and the lexicon. Hearing speech causes phonological representations to become active and enter the competition process described above. Each separate known form has a distinct stored representation; whenever we learn a new word, we have to construct a phonological representation and store it, if we want to recognize the word when we hear it again. Thus, many speakers of English acquired a new lexical representation when *bling* entered the language a few years ago. The process of learning a new phonological form and creating a corresponding lexical entry rarely causes problems in the native language, though adult foreign-language learners are aware that the process can be effortful. At issue in this section is the relationship between phonological representations in the lexicon and the prelexical representations computed during the processing of speech. It is easy to assume that the former will be entirely determined by the latter. If we hear two forms as distinct, and thus derive distinct prelexical representations for them, then their lexical forms will also be different; but if we hear them as indistinguishable they will

be assigned identical phonological representations prelexically, and hence the representations in the lexicon should also be identical. But it appears that this simple state of affairs does not hold. The phonological representations in the lexicon are abstract and can be codetermined by other sources of knowledge than our prelexical phonetic perceptions.

Second-language (L2) learning offers a useful window onto the relationship between prelexical and lexical processing. In the first language (L1), at least the very earliest stages of word learning in infancy must be fully dependent on the input; but in an L2, even the initial acquaintance with phonological form can be influenced by other sources of information as well as by the nature of the input. L2 learners already have an L1, for example; this means they have a set of phonemic categories that may help or hinder the formation of the correct set of categories for interpreting speech in the L2. They also have a good deal of knowledge about how words are structured phonologically, and, again, this can be helpful or unhelpful to the extent that the L1 and L2 do or do not match. An L1 that uses stress to distinguish words may hinder acquisition of an L2 that uses lexical tone; an L1 that does without articles before nouns may make acquisition of article use in an L2 very hard. Beyond the phonological mapping between the L1 and the L2, the adult learner brings many abstract expectations to the language-learning task: expectations about how different types of concepts are represented by nouns and verbs, for instance, or about how language can be used in different communicative situations. Finally, the L2 learner can receive explicit instruction, either formally from a teacher, or less formally, for example from workmates, and can draw on orthographic representations of new forms as well as spoken input. L2 learners exploit every type of help they can get with the language-learning task, and one result is that they set up phonological representations in the lexicon that include information that they have not extracted from the input.

This conclusion is supported by results from a series of L2 spoken-word recognition experiments, using a sensitive task which can examine the process of recognition moment by moment as the speech input constituting the target word unfolds. In this task, participants wear a head-mounted camera, by means of which the gaze direction of their eyes can be tracked (see Tanenhaus & Spivey-Knowlton, 1996). Typically, the participants are presented with a computer display containing several objects and are asked to carry out simple operations involving these objects, using the computer mouse. Their eye movements show that they actively consider alternative options for what the speech input could turn out to be and allow the incoming speech to modulate these options in a continuous manner. Thus if the display contained, for example, a camel, a corkscrew, a candle, and a frog, the very beginning of the frication noise for the initial /f/ of *frog* would be enough for listeners to select the frog, whereas the stop consonant with which the other three names begin would be enough to rule out the frog. If the stop consonant were followed by the vowel /æ/ of *camel* and *candle*, the corkscrew would then likewise be ruled out. But importantly, listeners do not just wait passively for the input to determine unambiguously what the target is—the task works so well for psycholinguistic purposes because alternatives that are still possible receive a significant proportion of looks, up to the moment at which the disambiguating further information arrives. Given the first two phonemes of *camel* and *candle*, both of the pictures will receive some looks until the nature of the following nasal phoneme supports one and disfavours the other. Thus the task allows us to observe phonetic processing for word recognition in a continuous manner; it allows us to see how the listener selects the correct word and what alternatives are considered (alternatives from the response selection set, of course, but potentially also alternatives in the vocabulary as a whole; Magnuson, Dixon, Tanenhaus, & Aslin, 2007).

Available candidates under active consideration as speech is heard can differ for L2 versus L1 listeners, if the L2 listeners' phonetic processing differs from the native norm. Presented with a

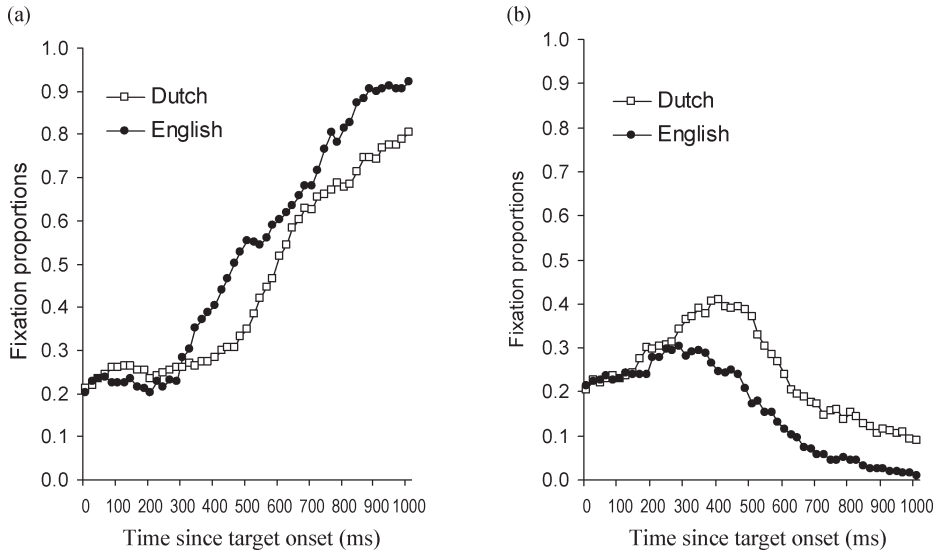

Figure 2. *Eye-tracking data for Dutch and British English listeners given spoken instructions such as* click on the panda *and a visual display containing inter alia a panda and a pencil; fixation proportions across time for (a) the target picture (panda), and (b) the competitor picture (pencil). Based on results from Weber and Cutler (2004).*

display containing a panda and a pencil and a duck and a strawberry, for example, and instructed in English to *click on the panda*, Dutch speakers of English as an L2 are very likely to look first at the pencil instead of at the panda (Weber & Cutler, 2004). This happens because Dutch listeners find it hard to distinguish the English vowel contrast exemplified in *pan-* versus *pen-*. Dutch has only one vowel in this area of the vowel space, and it is identical to neither of the English vowels but falls somewhere between them (though in broad phonetic transcription it is assigned the same representation as the vowel in English *pen*). Thus the Dutch category can capture both of the English vowels, making the distinction the hardest kind for an L2 learner to acquire (Best, 1995).

Figure 2 compares the looking pattern to (a) the target pictures (e.g., the panda) and (b) the competitor pictures (e.g., the pencil) for, in each case, Dutch participants and a control group of British English participants (the 10% or so of looks with which each group favoured the other two distractor items in the display are not shown in these figures). Across time, starting in these graphs from the onset of the spoken target word, the English participants look more rapidly to the correct target and away from the competitor, while the Dutch clearly look significantly more at the competitor item with a confusable vowel, and their looks to the correct target start to come in later.

This pattern would be consistent with an account in which the initial syllables of *panda* and *pencil* are totally indistinguishable in the Dutch listeners' prelexical phonetic processing of the input, and the stored phonological representations of the two words at the lexical level in consequence have identical initial syllables. This in turn would predict reversibility of the pattern. If the listeners simply cannot tell the difference, so that *pan-* and *pen-* are represented as homophonic in their lexicons just as *sale* and *sail* would have to be, then each form should be confused with the other. Instructed to *click on the pencil*, they should be quite likely to look at the panda initially instead.

But in fact this does not happen. Figure 3 shows the pattern that results when Dutch listeners hear *click on the pencil*—they obediently look at the pencil and do not appear to be tempted to look

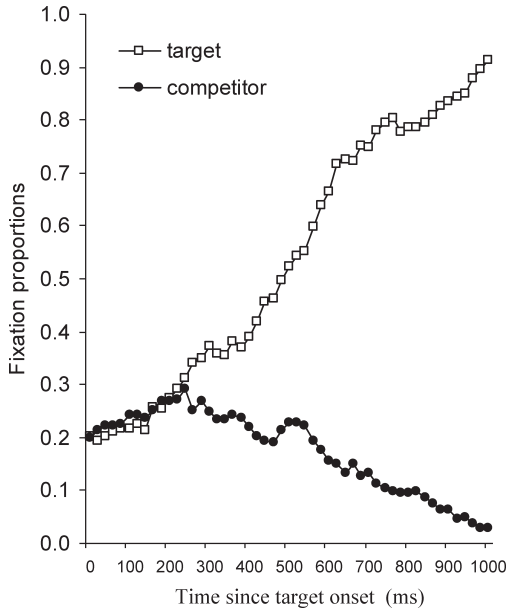

Figure 3. *Eye-tracking data for Dutch listeners given the same visual display as that in Figure 2, but spoken instructions such as* click on the pencil*; fixation proportions across time for the target picture (pencil) and the competitor picture (panda). Based on results from Weber and Cutler (2004).*

towards the picture of the panda at all. Thus the confusability they suffer from is asymmetrical; *pan-* could be either *pan-* or *pen-*, but *pen-* seems to be unambiguously *pen-*.

This happens not only with this vowel contrast for Dutch listeners to English, but also with the most widely studied second-language phonetic confusion, English /r/-/l/ for Japanese listeners. Results from a similar experiment by Cutler, Weber, and Otake (2006) are summarized in Figure 4. In Figure 4a, Japanese listeners instructed to click on a picture of a rocket experience noticeable extended competition (until at least 700 ms after target onset) from a picture of a locker, but in Figure 4b, listeners from the same group instructed to click on a picture of a locker look at it more rapidly and look less (from 400 ms after target onset) towards the alternative competitor picture (the rocket).

In both cases, then, the two L2 phoneme categories are not treated as equivalent; one of them is dominant. For the Dutch listeners, it is the vowel /ɛ/ as in *pen*; whether they hear *pan-* or *pen-*, they interpret it by preference as *pen-*. For the Japanese listeners, it is /l/; whether they hear *lock-* or *rock-*, they tend to interpret it as *lock-*. This is easily explicable in terms of closeness to the L1 category; the single Dutch vowel may fall between the two English vowels, but it is classified in the International Phonetic Alphabet (IPA) as /ɛ/, and English /ɛ/ is indeed a better match to it than English /æ/ is. Dutch listeners to English identify noise-masked syllables with /æ/ more often as /ɛ/ than vice versa, although English native listeners do not show such an asymmetry (Cutler, Weber, Smits, & Cooper, 2004). The single Japanese consonant that is closest to English /r/ and /l/ is represented by the letter *r* in transliterated words (e.g., *tempura*, *Hiroshima*, *harikiri*), but is articulatorily closer to /l/, being made with the tongue against the palate as /l/ is. When Japanese listeners are asked to rate English syllables on goodness of fit to Japanese syllables, they accord higher ratings to syllables with /l/ than to syllables with /r/ (Iverson et al., 2003; Takagi, 1995).

Thus in both cases there is a tendency for the percept to be classified as the dominant (most L1-like) category, whichever of the two L2 categories has actually been heard in the input. At the perceptual level, the L2 listeners are clearly not displaying accurate discrimination. At the lexical level, however, a quite different situation obtains: The way the undiscriminated percept is mapped to the lexicon is surprisingly veridical. The input that was categorized by Dutch listeners as /ɛ/, or the input that was categorized by Japanese listeners as /l/, indeed contacts the lexical entries that are supposed to contain these phonemes (*pencil*, *locker*) and not the entries that are supposed to contain the other members of these phonemic contrasts (e.g., *panda*, *rocket*). Why is this surprising? Of course, it would not be at all surprising for a native listener to achieve the same—hear *pen-*, and map it to *pen* or *pencil* and absolutely not to *pan* or *panda*, or hear *lock-* and map it to *lock* or *locker* and absolutely not to *rock* or *rocket*. But how have the L2 listeners managed to do it? If the input tends to be

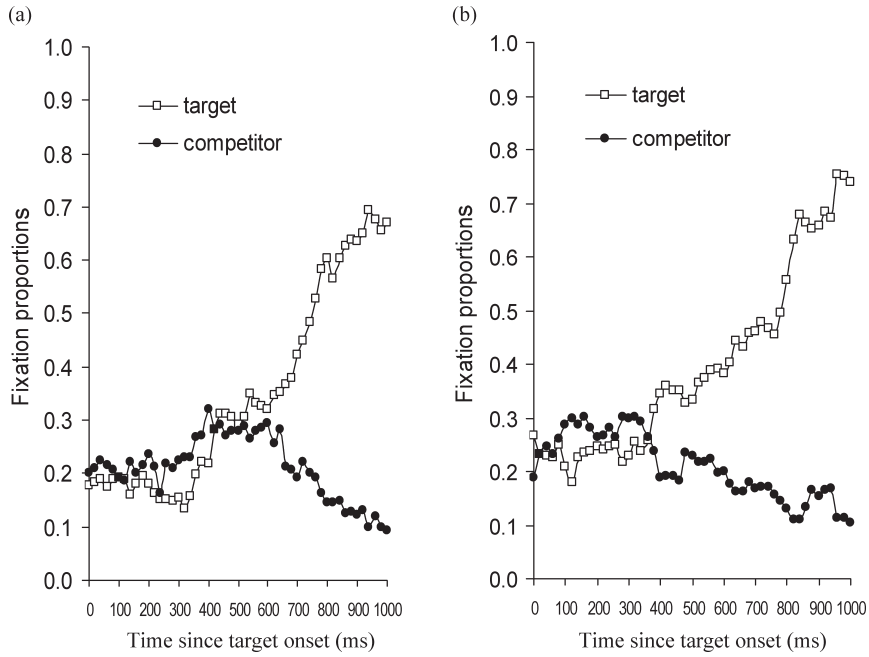

Figure 4. *Eye-tracking data (fixation proportions across time for target picture and competitor picture) for Japanese listeners given a visual display containing inter alia a locker and a rocket; (a) for spoken instructions such as* click on the rocket *(target = rocket, competitor = locker), (b) for spoken instructions such as* click on the locker *(target = locker, competitor = rocket). Based on results from Cutler, Weber, and Otake (2006).*

interpreted as the same thing whichever phoneme is said (i.e., the listeners cannot hear the difference), how do they know to map the input to one lexical entry and not to the other?

Clearly, the phonological representations for /æ/ and /r/ in these listeners' lexicons are not identical to the representations for /ɛ/ and /l/, and thus they are not contacted by the access representations that successfully contact the lexical representations with /ɛ/ and /l/. Phonological representations in the lexicon therefore cannot be determined solely by the listeners' experience with spoken input. In the case of the nondominant member of the L2 pair (/æ/ for Dutch listeners, /r/ for Japanese listeners), the lexical listing does not accurately reflect the categorization that results from exposure to the sound in question in a spoken word. Instead, it reflects a separate categorization, which, however, is one that the listeners usually do not come up with. The process of contacting the lexical representations of these words is thus an interesting topic for study, and varying interpretations of the process are possible (see Cutler et al., 2006, for discussion of this point).

Another interesting issue is the range of potential factors that can modulate the information from auditory perception in establishing phonological representations in the lexicon. As noted above, L2 learners have a variety of knowledge sources to draw on, beyond the information they receive in the spoken forms of words. At least explicit pronunciation instruction, orthographic representation, and form familiarity would seem to offer potential information about whether stored phonological representations should be the same or different, and possibly L2 learners draw on more than one such source. In a clever follow-up to Weber and Cutler's (2004) study, Escudero, Hayes-Harb, and Mitterer (2008) taught Dutch speakers of English the supposed English names for some nonsense figures, including pairs like

tanzer and *tendik*, which differed in much the same way as *panda* and *pencil*. When the nonsense names were taught using only the spoken forms, the two were confused equally often with each other in a subsequent test with eye-tracking; the instruction *look at the tan-* produced about the same proportion of looks to the tanzer as to the tendik, and so did the instruction *look at the ten-*. These listeners could not tell the vowels apart and treated the two initial syllables as versions of the same syllable. When the nonsense names were taught using the spelled as well as the spoken form, however, the eye-tracking results were quite different: The instruction *look at the tan-* produced looks to the tendik as well as the tanzer, but the instruction *look at the ten-* led the listeners to look overwhelmingly at the tendik only. This is exactly the same result as Weber and Cutler had observed. Thus the additional orthographic information enabled these listeners to realize that the vowels were supposed to be different and hence to incorporate differences in the first syllables in their stored representations of the novel words, in just the same way as they did for known English words such as *pan* and *panda* versus *pen* and *pencil*. But just as with real English input, the availability of a lexical distinction did not suffice to induce a corresponding sensitivity in the listeners' prelexical phonetic processing.

The lesson to be drawn for present purposes concerns the nondeterministic relationship of phonological processing at the prelexical level and phonological representations at the lexical level; the two are, at least to a large extent, independently determined. Listeners use abstract knowledge about phonological distinctions, derived for instance from orthography, to shape the stored phonological forms; once they know there is supposed to be a distinction between two phonemes, they store the distinction, even if they can't reliably hear it. This is very easy to see in the L2 case, because the phonemic confusions involved are predictable from the L1–L2 mapping. But there is no reason in principle why the independence should not be equally great in the L1 case. It is only in infancy, in the very earliest stages of word learning, that we are necessarily confined to using only spoken traces of the word forms; as soon as we can interact with others, we are open to explicit instruction and hence to the exploitation of other sources of information about words. Once we learn to read, the possibilities expand further. Many words that educated adults know have been learned from reading and may indeed never have been heard (sometimes even with the result that the stored phonological form is quite incorrect). The lexical representations that language users draw on are richly supplied with phonological information, only some of which has been drawn from prelexical processing of speech input.

The necessity of abstract prelexical representations

Now consider the nature of this prelexical processing and the representations it involves. As noted in the introduction, models of speech processing traditionally assumed that these representations were relatively abstract. A process of normalization rids speech input of all its speaker- and situation-specific properties, leaving over only the communicative essence in some phonetic form. A range of phenomena has been ascribed to such abstraction, from interpretation of variant forms of words (*postman* versus *pos'man*, *film* versus *fillum*) as the same canonical lexical form (Donselaar, Kuijpers, & Cutler, 1999; Sumner & Samuel, 2005) to different effects of phoneme transition probability on the processing of spoken words and nonwords (Vitevitch, 2003; Vitevitch & Luce, 1998). The ease with which listeners can deal with speech from talkers whose voices they have never previously heard has always been one of the strongest motivations for the classical model. Indeed, our subjective experience is that understanding an utterance from a new talker—for instance, when a stranger in the street asks for directions, or a shopkeeper names a price—is usually no harder than understanding the same utterance from a speaker whose voice is familiar to us.

However, recent evidence for talker-familiarity effects in spoken-word recognition has motivated

some quite different approaches. It appears that hearing a word spoken a second time by the same speaker can result in easier recognition than hearing a word spoken a second time but by a different speaker—at least, if the task is to decide whether the word had been heard before (Goldinger, 1996; Luce & Lyons, 1998) or to identify words against a noisy background (Mullennix, Pisoni, & Martin, 1989; Nygaard, Sommers, & Pisoni, 1994). Such evidence prompted the development of models in which traces of every speech input experience are stored, and in some of these models the necessity of any abstraction at all is queried. This radical step is not a good idea, though; there is abundant evidence that spoken-language recognition necessarily involves abstract prelexical representations, separate from the representations of words. This section describes some of this evidence.

Perceptual learning effects in phoneme perception

Among the speaker-specific properties of speech signals is variation in phoneme realization. Some speakers produce particular phonemes in an unusual way—for example, with a lisp, or a foreign accent. Experience tells us that unusual phonemic realizations are sometimes initially problematic, but adaptation to a deviant pronunciation can be extremely rapid. Norris, McQueen, and Cutler (2003) developed a two-stage procedure for inducing such adaptation. Participants in their study first took part in an auditory lexical decision experiment. A total of 20 words among the items of this experiment contained a phoneme the pronunciation of which had been manipulated, so that it fell approximately halfway between the speaker's natural /s/ and natural /f/. The phoneme was spliced into words where it replaced either /s/ or /f/. For some participants, /f/ was replaced by the ambiguous phoneme; so they heard 20 words like *loaf* or *enough* or *handkerchief*, each ending with the deviant phoneme instead of a natural /f/ (note that although these are English examples, the experiment was actually in Dutch). They also heard 20 words with a clear /s/ in final position, such as *lace* or *carcase* or *numerous*, and 160 other words and nonwords none of which contained any /f/ or /s/. Other participants heard the 20 /s/-words (*lace*, etc.) ending with the deviant phoneme, while the /f/-words (*loaf*, etc.) were heard in their natural pronunciation; otherwise, the lexical decision experiment was the same for both these groups. The first group of participants thus should learn that this speaker pronounces /f/ in an odd way, while the second group should learn that it is /s/ that it is oddly pronounced; in either case, the listeners should adapt to the deviant pronunciation.

They certainly did, as the results of the second part of the experiment showed. In the second part, the listeners performed a phonetic categorization task, in which they heard vowel–consonant syllables and decided whether the consonant was /f/ or /s/. The consonant varied along a continuum from a form close to a good /f/ to a form close to a good /s/. Typically responses in such an experiment show a categorical function – in the early part of the continuum the percentage of choices for one category is high, then there is a rapid cross-over to a low percentage for that category in the later part. The function for continuous sounds such as vowels or, in this case, fricatives, is less categorical (i.e., has a less steep cross-over) than the function for discontinuous sounds such as stop consonants, but still shows a clear progression from one category choice to the other. A shift in the boundary between the two categories typically shifts the function to the left or the right in the middle range of the continuum (but does not affect the end points, which are still clearly assigned to the one or the other category).

Norris et al.'s (2003) phonetic categorization test was quite short and omitted the end points; it included five ambiguous sounds varying in closeness to /f/ and /s/. Results from their Experiment 2 are shown in Figure 5. The two middle functions are results from two control groups of subjects, who had heard nonwords containing the same ambiguous sound as the groups described above had heard in real words. One control group had heard the ambiguous nonwords plus the /s/-words such as *lace*, while the other control group had heard the ambiguous nonwords plus the /f/-

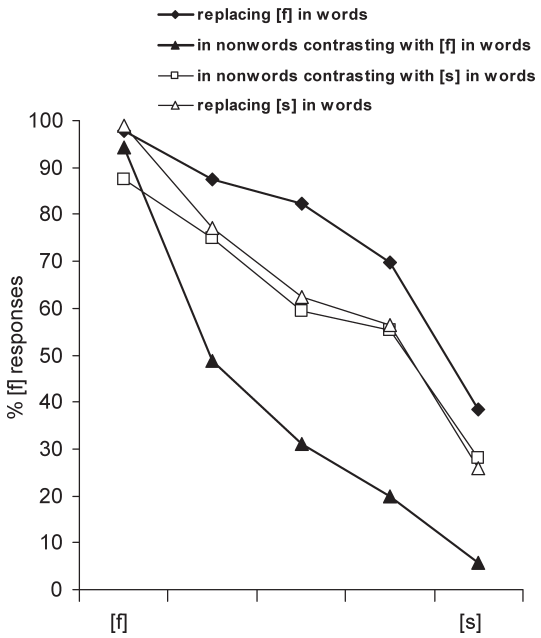

Figure 5. *Phoneme categorization data after perceptual learning for an ambiguous phoneme realization: Proportion of /f/ responses after hearing the ambiguous sound replacing /f/ in words (as in* loaf*), replacing /s/ in words (as in* lace*), in nonwords with contrasting words containing /s/, or in nonwords with contrasting words containing /f/. Based on results from Norris et al. (2003).*

words such as *loaf*. As Figure 5 showed, this difference did not significantly affect their responses. But the two experimental groups who had heard the ambiguous sound in the context of real words certainly showed a significant difference in their response patterns. The group that had learned that this sound was to be interpreted as /f/ produced more /f/ responses across most of the continuum; the group that had learned that this sound was to be interpreted as /s/ produced fewer /f/ responses. Thus their responses were affected not just for the single ambiguous sound heard in the first phase, but across a range of the continuum, consistent with a shift in the boundary between their /f/ and /s/ categories. The learning concerned not a specific acoustic form, but the relation between these abstract categories. The direction in which the boundary shifted was then different for each group, consistent with adaptation to the differing identities that they had been trained to allot to the ambiguous sound.

In Norris et al.'s (2003) study, the appearance of the adaptation was dependent upon the availability of lexically provided information about how the ambiguous sound should be interpreted. Listeners did not alter their phonetic categorizations after being exposed to a deviant pronunciation in nonwords, even though they could perhaps have deduced the putative identity of the strange sound from distributional information in the remaining items (e.g., the clear /s/ in words like *lace* would suggest that the sound was not /s/). The lexical training, however, was extremely effective and notably rapid; hearing a deviant realization of one phoneme in just 20 words was enough to induce listeners to shift the boundaries between that phoneme and a neighbour. Later research showed that even then, Norris et al. had overestimated the number of training exposures necessary; in a study by Kraljic and Samuel (2006) 10 exposure words proved sufficient to induce a significant shift. Lexical decision, which demands attention to the word's form to enable the decision that, say, *word* is a word but *worb* is not, is also not a necessary precondition for induction of the effect; when the words occurred in a story (Eisner & McQueen, 2006), the shift was just as robust. The phonemes in that story occurred in different positions in the words, so a constant position was also unnecessary, and even if position in the word is kept constant during training, the perceptual learning can transfer at test to occurrences of the same phoneme in a different position (Jesse & McQueen, 2007). Further, induction of a shift was also just as robust if the words did not have to be processed as words at all, but listeners merely tallied how many words they heard (McQueen, Norris, & Cutler, 2006c). Lexical information is not the only type of information that listeners can use for such learning—visual speech cues will also do the trick (Bertelson, Vroomen, & de Gelder, 2003).

Perceptual learning as speaker adaptation

If the way that we adapt speech perception to individual speakers involves perceptual learning of this

kind, then a few entailments obviously follow. First, the adaptation that appears in a learning experiment of the sort just described should be confined to the case of speech from a particular speaker, speaking a particular dialect. It would not be helpful to generalize too widely. After all, if we are chatting with a group of Scottish friends, and a speaker with a Birmingham accent joins the conversation, adaptation to Brummy phoneme settings will help for input from that speaker, but is likely to be counterproductive if applied when the Scots are talking. Eisner and McQueen (2005) tested the speaker specificity of the adaptation by training listeners with lexical decision items spoken by one talker and testing with a phonetic continuum from another talker. There was no adaptation in that case. If, however, the fricative continuum came from the same talker even though the participants were unaware of this (because it had been inserted into speech from another talker), the adaptation was observed. Note that fricatives encode speaker-specific information quite effectively, so listeners can easily restrict the adaptation to a specific talker's speech. Kraljic and Samuel (2007) showed that it is possible to train listeners to construct speaker-specific fricative representations concurrently for more than one talker in these experiments. In contrast to fricatives, stop consonants do not vary as much across speakers, and Kraljic and Samuel (2006, 2007) found that adaptation for word-medial /d/ or /t/ (e.g., in *academic*, *cafeteria*) did generalize across speakers, albeit with an effect size much smaller than the effect size that they had observed for fricatives.

Second, it is clearly a prerequisite of adaptation to talkers that it should last. Ideally, knowledge about characteristic articulatory idiosyncrasy should remain helpful the next time we meet the same person. This is also indeed the case. Kraljic and Samuel (2005) found that the perceptual learning lasts for 25 minutes in a laboratory testing session filled with other tasks; Eisner and McQueen (2006) found that it lasts for 12 hours, irrespective of whether those hours were mainly filled with sleep or were daytime hours filled with normal daytime activities. What can effectively and quickly reduce the adaptation is hearing the same talker produce nondeviant renditions of the same phoneme. Clear productions by another talker have no effect, but a clear version by the person responsible for the apparently deviant production abolishes the phoneme boundary shift right away (Kraljic & Samuel, 2005). This too is just as required: If a friend has a cold, we want to be able to understand the modified speech, but we want to be able to adapt back right away once the nasal congestion has passed. Note that if listeners acquire evidence that a deviant articulation is not due to speaker-specific idiosyncrasy, but is a dialectal feature restricted to a single phonetic context (Kraljic, Brennan, & Samuel, 2008a), or is caused by a transient state such as speaking with a pen in the mouth (Kraljic, Samuel, & Brennan, 2008b), they do not adjust phoneme boundaries; the prerequisite of adaptation is that the deviant pronunciation is indeed likely to be encountered in the same talker's future utterances.

Third, and most importantly, the adaptation should generalize beyond the lexical items involved in the training situation. Otherwise it would be useless. Our experience is that when we adapt to a new talker, we understand whatever that talker says even though we have no stored memories of those particular words from that talker. So hearing someone pronounce /f/ in an odd way in *loaf*, *carafe*, and *handkerchief* should help us understand the same person's later pronunciations of, say, *dwarf* or *hoof*. Likewise, it should enable us to distinguish between potential minimal pairs—for instance, to know that an utterance of *knife* is not *nice*. This would obviously not occur if learning is specific to the particular utterances that have been heard.

Generalization to other words would imply that, as argued above, the learning is happening at a truly abstract level; it concerns phonemic categories, independently of the forms in which they might be encountered.

McQueen, Cutler, and Norris (2006a) have established that generalization to other words does indeed occur. They combined adaptation,

induced using the training conditions of Norris et al. (2003), with a lexical test phase to assess whether the adaptation had generalized. The training involved, again, an f/s continuum, with half the subjects learning that the ambiguous sound was /f/ and the other half learning that it was /s/; the test phase made use of minimal pairs of the *knife*/*nice* kind. The same utterance, ending with the same ambiguous sound, should be recognized as an utterance of *knife* by the listeners who had been trained that the ambiguous sound was /f/, but it should be recognized as *nice* by the other group, who had been trained to identify the ambiguous sound as /s/. To determine which word the listeners had recognized, McQueen et al. employed the cross-modal identity priming paradigm, which, as we saw above, tests for activation of a phonological representation in the lexicon. A spoken prime that was a potential minimal pair was presented, with the ambiguous sound replacing its final phoneme. If such a prime is recognized as *knife*, then it will facilitate responses to visually presented KNIFE (but will not facilitate NICE). This is the pattern we would predict for subjects with /f/-biased training. For subjects with /s/-biased training we would predict the opposite: The prime should be recognized as *nice*, and it will facilitate responses to visually presented NICE but not KNIFE.

Figure 6 summarizes the results of their study. The priming effect when the target word was consistent with the interpretation of the prime induced by the training (KNIFE given /f/ training, NICE given /s/ training) is compared with the priming effect when the target word was inconsistent with the training-induced interpretation (KNIFE given /s/ training, NICE given /f/ training). Clearly, responses to targets consistent with the interpretation are significantly facilitated, while responses to inconsistent targets are not facilitated at all.

Thus the shift in a phonemic category's boundary induced by a short training session involving 20 words has the desired consequences: It generalizes more or less immediately across the lexicon such that perception of any word involving that

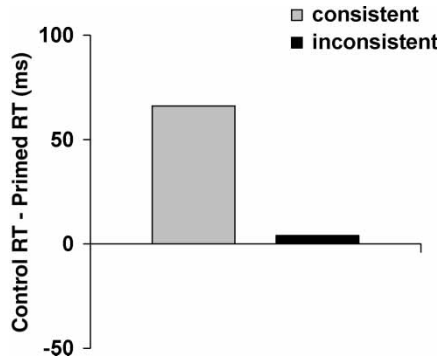

Figure 6. *Cross-modal identity priming effects (difference from control prime condition) of spoken prime words ambiguous between minimal pairs such as* knife/nice*, for visual target words consistent with the trained interpretation of the ambiguous final phoneme (KNIFE after /f/ training, NICE after /s/ training) versus inconsistent with it (KNIFE after /s/ training, NICE after /f/ training). Based on results from McQueen et al. (2006a).*

phoneme is appropriately adjusted. This generalization has since been replicated, with a different ambiguous sound in the training condition, by Sjerps (2007), who also established that the strength of the posttraining priming by words with the ambiguous sound was indistinguishable from that of the priming by naturally spoken tokens of the same words.

The generalization, to words in which the trained sound occurs in new phonetic contexts, implies that the learning has occurred at a level involving phonemic representations—abstract forms standing for the sound that is common to *carafe*, *handkerchief*, and *knife*, and indeed to *fate*, *fresh*, *office*, and *muffle* as well. A model that denies all abstraction and confines stored representations to the traces of already-experienced episodes cannot account for this finding. MINERVA-2 (Hintzman, 1986), as used by Goldinger (1998) in his model of lexical access, is such a model; Cutler, Eisner, McQueen, and Norris (in press) attempted to model the perceptual learning and generalization in an implementation of this model, but it in fact predicted the opposite pattern of results from that revealed in the McQueen et al. (2006a) experiment.

Conclusion

The new findings summarized here constrain the structure of speech-processing models. Human listeners make use of separate representations of incoming speech at a prelexical level and at a lexical level. The prelexical representations are abstract, requiring matching abstraction in the lexicon. Within the lexicon, the representations of word form and word meaning dissociate. The form representations in the lexicon are not fully determined by phonetic processing of speech input, but are also shaped by abstract knowledge of contrastive distinctions.

Some of these constraints (dissociable form and meaning representations in the lexicon; the independence of lexical form representations from prelexical processing) are novel, while others (abstraction; the separation of prelexical and lexical processing) are features of what has been referred to in preceding sections as "traditional" or "classical" psycholinguistic approaches to speech perception. Both abstraction and prelexical/lexical separation have been called into question by approaches based on episodic memory (e.g., Goldinger, 1998), but without them, such approaches are inadequate. (This is not news to the proponents of episodic models. Researchers with considerable episodic track records have recently called for hybrid approaches combining abstract and episodic information: Goldinger, 2007; Pisoni & Levy, 2007.) In any case, it is clear that models that deny a role for abstraction at any or all of the levels of human speech perception distinguished here, or which fail to acknowledge the separation of these processing levels, must be discarded.

The functional significance of abstract representations is substantial; abstraction and generalization play a major role in the efficiency of cognitive processing in general and speech processing in particular. Without abstraction, it would not be possible to adapt quickly to newly encountered talkers; communication would be a much slower and more errorful affair. Note that the capacity to retune perceptual category decisions is not specific to language—analogous retuning occurs, for instance, for colour categories (Mitterer & De Ruiter, 2008). Nor, within language, is it specific to speech—it operates in just the same way for an ambiguous letter in print (e.g., a letter that could be H or N, but is disambiguated by appearing at the end of *WEIG-* or of *REIG-*; Norris, Butterfield, McQueen, & Cutler, 2006a). Within speech, the knowledge that is drawn upon in interpreting an ambiguous sound need not be lexical; it can be, for example, constraints on permissible phoneme sequences, processed in nonwords (e.g., a sound that could be /f/ or /s/ is interpreted as /f/ if it precedes *-rumic* and as /s/ if it precedes *-nuter*, because the sequences /sr/ and /fn/ are illegal in English; Cutler, McQueen, Butterfield, & Norris, 2008). Decisions about abstract category membership permeate all domains of cognitive processing, and these similar effects across multiple domains suggest the involvement of a powerful general learning mechanism aimed at improving the efficiency of such category decisions by rapid reference to whatever meaningful knowledge is available.

There is much more to be said about the complete account of human speech processing. Integrating abstract representation and stored episodic information in a single model is a challenge, and there is as yet little directly relevant evidence (though see, for example, McLennan, Luce, & Charles-Luce, 2003, for an argument that while episodic representations may be drawn upon in easier speech-processing tasks, more difficult tasks require the availability of abstract representations). But this is far from the only challenge. The evidence for abstract representation must also be integrated with the evidence for continuous cascade of graded information across the levels at which speech is processed (see McQueen, Dahan, & Cutler, 2003, for a review). Here the future may lie with probabilistic accounts such as the Bayesian model of spoken-word recognition recently proposed by Norris and McQueen (2008). Further, it is characteristic of human listeners (though presumably not of dogs or monkeys) that the nature and course of operations at each level of processing is shaped

by the mother tongue. Listeners' phonemic representations are determined by what phonemes contrast in the native language, and their lexical representations by what words are found in the native vocabulary; but even beyond this, there are processing effects at every level. Thus the size and make-up of the phoneme inventory can cause differences in how one and the same contrast between two speech sounds is processed across languages; although /f/ and /s/ (as in *leaf–lease*) are articulated in the same way in Spanish, English, Dutch, and Italian, listeners attend to transitional cues in a preceding vowel to distinguish /f/ from /s/ in Spanish and in English, but not in Dutch or Italian, the reason being that the former two languages have the additional similar sound /θ/ (as in *teeth*), but the latter two do not (Wagner, Ernestus, & Cutler, 2006). Likewise, the distribution of sounds in the phoneme inventory determines how sensitive listeners are to one and the same source of variation; in languages with large numbers of both vowels and consonants, such as English, French, or Dutch, listeners are equally sensitive to effects of vowel variation on consonant realization and vice versa, but in languages with a highly asymmetric inventory, such as Spanish with four times as many consonants as vowels, listeners are far more sensitive to consonantal effects on vowels than the reverse (Costa, Cutler, & Sebastián-Gallés, 1998). The structure of the lexicon also influences whether or not listeners pay attention to one and the same source of acoustic evidence, such as suprasegmental cues to stress, which are attended to in the activation of words in Dutch (Donselaar, Koster, & Cutler, 2005) and Spanish (Soto-Faraco, Sebastián-Gallés, & Cutler, 2001), but are largely disregarded in English (Cooper, Cutler, & Wales, 2002). A complete account of speech processing will thus need to map the range of language-specific modulations of the language-independent substrate. The representations are, therefore, just one part of the story; however, they are an important part, and in constraining our account of the representations involved, we certainly come closer to a fuller understanding of one of humankind's favourite cognitive operations: listening to speech.

Original manuscript received 30 January 2008
Accepted revision received 2 May 2008
First published online 31 July 2008

REFERENCES

Arnold, K., & Zuberbühler, K. (2006). Language evolution: Semantic combinations in primate calls. *Nature*, *441*, 303.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*, 592–597.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (pp. 167–200). Timonium, MD: York Press.

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech*, *45*, 207–228.

Costa, A., Cutler, A., & Sebastián-Gallés, N. (1998). Effects of phoneme repertoire on phoneme decision. *Perception & Psychophysics*, *60*, 1022–1031.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (in press). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron (Ed.), *Papers in laboratory phonology 10*. Berlin, Germany: Mouton de Gruyter.

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. In J. Fletcher, D. Loakes, M. Wagner, & R. Goecke (Eds.), *Proceedings of Interspeech 2008* [CD-ROM]. Brisbane, Australia.

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*, 269–284.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, *116*, 3668–3678.

Dell, G. S. (1988). The retrieval of phonological forms in production: Test of predictions from a connectionist model. *Journal of Memory and Language*, *27*, 124–142.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*, 313–349.

Donselaar, W. van, Koster, M., & Cutler, A. (2005). Exploring the role of lexical stress in lexical recognition. *Quarterly Journal of Experimental Psychology*, *58A*, 251–273.

Donselaar, W. van, Kuijpers, C., & Cutler, A. (1999). Facilitatory effects of vowel epenthesis on word processing in Dutch. *Journal of Memory and Language*, *41*, 59–77.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*, 224–238.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America*, *119*, 1950–1953.

Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, *36*, 345–360.

Gaskell, M., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613–656.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)* (pp. 49–54). Dudweiler, Germany: Pirrot.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Ketterman, A., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*, B47–B57.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 824–843.

Jesse, A., & McQueen, J. M. (2007). Prelexical adjustments to speaker idiosyncracies: Are they position-specific? In H. van Hamme & R. van Son (Eds.), *Proceedings of Interspeech 2007*, Antwerpen, Belgium (pp. 1597–1600) [DVD]. Adelaide, Australia: Causal Productions.

Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: Evidence for fast mapping. *Science*, *304*, 1682–1683.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008a). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*, 54–81.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, *13*, 262–268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*, 1–15.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008b). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, *19*, 332–338.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–38.

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, *26*, 708–715.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*, 1–24.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*, 363–369.

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003) Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 539–553.

McQueen, J. M., Cutler, A., & Norris, D. (2006a). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113–1126.

McQueen, J. M., Dahan, D., & Cutler, A. (2003). Continuity and gradedness in speech processing. In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 39–78). Berlin, Germany: Mouton de Gruyter.

McQueen, J. M., Norris, D., & Cutler, A. (2006b). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*, 533.

McQueen, J. M., Norris, D., & Cutler, A. (2006c). The dynamic nature of speech perception. *Language and Speech*, *49*, 101–112.

Mitterer, H., & De Ruiter, J. P. (2008). Recalibrating color categories using world knowledge. *Psychological Science*, *19*, 629–634.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*, 365–378.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

Norris, D., Butterfield, S., McQueen, J. M., & Cutler, A. (2006a). Lexically-guided retuning of letter perception. *Quarterly Journal of Experimental Psychology*, *59*, 1505–1515.

Norris, D., Cutler, A., McQueen, J. M., & Butterfield, S. (2006b). Phonological and conceptual activation in speech comprehension. *Cognitive Psychology*, *53*, 146–193.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357–395.

Norris, D., McQueen, J. M, & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42–46.

Pisoni, D. B., & Levi, S. V. (2007). Some observations on representations and representational specificity in speech perception and spoken word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 3–18). Oxford, UK: Oxford University Press.

Sjerps, M. J. (2007). *Nonnative phonemes are open to native interpretation: A study on the flexibility of speech perception*. Unpublished MSc thesis, Radboud University, Nijmegen, The Netherlands.

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language*, *45*, 412–432.

Sumner, M., & Samuel, A. G. (2005). Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language*, *52*, 322–338.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning & Verbal Behavior*, *18*, 645–659.

Takagi, N. (1995). Signal detection modeling of Japanese listeners' /r/-/l/ labeling behavior in a one-interval identification task. *Journal of the Acoustical Society of America*, *97*, 563–574.

Tanenhaus, M., & Spivey-Knowlton, M. (1996). Eye-tracking. *Language and Cognitive Processes*, *11*, 583–588.

Vitevitch, M. S. (2003). The influence of sublexical and lexical representations on the processing of spoken words in English. *Clinical Linguistics & Phonetics*, *17*, 487–499.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in the perception of spoken words. *Psychological Science*, *9*, 325–329.

Wagner, A., Ernestus, M., & Cutler, A. (2006). Formant transitions in fricative identification: The role of native fricative inventory. *Journal of the Acoustical Society of America*, *120*, 2267–2277.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*, 1–25.